A SVM and k-NN Restricted Stacking to Improve Land Use and Land Cover Classification

Jorge Garcia-Gutierrez, Daniel Mateos-Garcia, and Jose C. Riquelme-Santos

Department of Computer Science
E.T.S.I.I. - University of Seville, Spain
{jgarcia,mateos,riquelme}@lsi.us.es

Abstract. Land use and land cover (LULC) maps are remote sensing products that are used to classify areas into different landscapes. The newest techniques have been applied to improve the final LULC classification and most of them are based on SVM classifiers. In this paper, a new method based on a multiple classifiers ensemble to improve LULC map accuracy is shown. The method builds a statistical raster from LIDAR and image fusion data following a pixel-oriented strategy. Then, the pixels from a training area are used to build a SVM and k-NN restricted stacking taking into account the special characteristics of spatial data. A comparison between a SVM and the restricted stacking is carried out. The results of the tests show that our approach improves the results in the context of the real data from a riparian area of Huelva (Spain).

1 Introduction

Remote sensing has become a very important tool to carry out many different tasks for the Natural Environment. In this way, remote sensing has successfully been applied to important activities like flood control, forestal inventories or invasive species control in protected or specially interesting areas.

Although remote sensing usually works with images exclusively, data fusion has been of high interest since the appearance of new active sensors (i.e., data is produced as a response for a stimulus which is not the solar light). They complement images and overcome some of their limitations, e.g., the problems associated to shadows. These limitations cause fusion of sensors can be found as a proper technique specially interesting to improve the results of the classical remote sensing approaches. One of the most active research lines has been based on LIDAR (LIght Detection And Ranging) technology. This technology is able to register object heights and it is specially recommended to be applied on complex landscapes like riparian zones. Thus, Verrelst et al.[1] use LIDAR to study vegetal species communities and Antonorakis et al.[2] develop a new methodology to identify different types of commercial wood in riparian zones using only LIDAR.

An automatic pixel classification which is generally supervised is usually the first step to extract knowledge from remote sensing data. Several techniques from machine learning have been used with satisfactory results though support vector machines (SVM) are the predominant technique to obtain the best results

E.S. Corchado Rodriguez et al. (Eds.): HAIS 2010, Part II, LNAI 6077, pp. 493–500, 2010.

in most cases [3]. Despite the SVM's high accuracy, improvement is needed to reach the standards for products like land uses and land cover (LULC) maps [4].

LULC maps are remote sensing products that are used to classify areas into different landscapes subject to their own characteristics or functionality. The newest techniques have been applied to improve the final classification to develop LULC maps. Fauvel et al. [5] apply an SVM to classify the pixels depending on morphologic and hyperspectral data. In Mitrakis et al. [6], a neuronal network with weights determined by a genetic algorithm obtains the final classification using fusion operators and fuzzy logic. It is important to underline that ensembles are one of the most powerful tools in machine learning and so they are in remote sensing where they have also been applied profusely. A very clear example can be seen in [7] where an stacking of several SVM's and a random forest is used to carry out the pixel classification.

This work explores the application of ensembles on remote sensing taking advantage of contextual information [8] from multi-source (LIDAR and aerial images) data. Thus, a novel supervised method called R-STACK (based on a stacking of a SVM and multiple NN classifiers) is shown with two purposes:

- Show an easy way to improve the quality of models when intelligent techniques are applied on LIDAR and imagery fusion data.
- Improve the general accuracy of an automatically generated LULC map.

The rest of the paper is organized as follows. Section 2 provides a description of the data used in this work. Section 3 describes the methodology used, highlighting the feature set and the model extraction process. The results achieved are shown in section 4 and, finally, section 5 is devoted to summary the conclusions and to discuss future lines of work.

2 Data Description

A LIDAR system is an optical sensor technology that measures properties of scattered light (usually laser) to find range and/or other information of a distant target. The whole process starts with the emission of polarized light, typically, in the ultraviolet visible or near infrared. Then, LIDAR catches the reflected signal from the topographic surface and measures the time employed for each return to establish the distance between the emitter and the object that produced the return. This process is helped by a global positioning system (GPS) to give rise to a cloud point database in which for every point, it is possible to find: spatial position(i.e., x, y and z coordinates), intensity of return, number of the return in a sequence (if a pulse caused multiple impacts), etc. This features and the RGB values in an orthophoto are used in this work to obtain statistical measures on which the method is based and they will be explained in section 3.

The LIDAR data was taken in coastal areas of the province of Huelva. The pulses were geo-referenced and correctly validated by the distributor of the data and having 1,384,875 records for an area of 1.5 km^2. The reported precision indicates a maximum error of 0.5 m in the x-y positions, and of 0.15 m in the

z position. Along with the LIDAR flight, the aerial photographs were taken of the area with a resolution of 0.5 m^2.

The study area is situated in the south of Spain, in the mouth of the Tinto and Odiel rivers. This area is near the city of Huelva and presents a mix of urban and natural areas. The natural areas can be classified in five subclasses: watered zones, marshland and vegetation (low, middle and high). The high vegetation is formed by scarce trees of the genus *eucalyptus* in the area. The middle vegetation is formed by different types of Mediterranean bushes that principally surround roads and urban areas. Pastures are classified as low vegetation and include bare earth areas. In addition, the urban areas are also classified in five subclasses: roads and railways, buildings, coal deposits, dumps and mixed areas.

3 Method

Our LULC development method (see Figure 3) follows a pixel-oriented strategy which obliges us to create a matrix or raster where each element is a pixel. Each pixel represents an area in function of the resolution. The value of resolution must be provided by the user as a method parameter to determine the area within each pixel. The resolution depends on the LIDAR point density and the orthophoto resolution. In our case, the selected resolution is set at 3 m^2. Lesser resolution could damage the smallest classes (roads) classification and bigger resolution cannot be possible due to the LIDAR resolution (0.5 points/m^2).

Apart from the resolution, it is necessary to supply a digital elevation model (DEM) to extract the actual heights of the LIDAR returns. In our case, this process is carried out by an adaptative morphologic filter [9]. In addition, expert knowledge has been applied to manually classify over a 2% of total data (7399 instances). Expert knowledge leaned on the photographs taken in the same flight as LIDAR data and previous information from the Regional Ministry of Andalusia (LULC map from 2003) was collected by an operator to build the training set.

The next step (step 2 in Figure 3) is to calculate a set of variables from image RGB values, LIDAR intensity, heights and distribution of the LIDAR returns for each pixel (a total of 500,000 pixels). In this manner, sixty-one different measures were calculated for every pixel. Most of variables used have been extracted from literature [10][2]. In Table 1, a summary of these features can be seen. Specially interesting is the case of the normalized difference vegetation index (NDVI). The classical NDVI is generated from near infrared band (NIR) and the red band (R) as can be seen in Equation 1. In our case, it cannot be calculated since the NIR band is not available in LIDAR or orthophotography. Thus, the new attribute SNDVI has been used to simulate the NDVI using the intensity (I) from LIDAR (Equation 2) as near-infrared value which approximates the real NIR value.

$$NDVI = \frac{NIR - R}{NIR + R} \tag{1}$$

$$SNDVI = \frac{I - R}{I + R} \tag{2}$$

A new method called R-STACK based on a modified stacking of two well-known classifiers (SVM and k-NN) has been developed for the model generation. The Weka [11] implementation of SOM and a ad-hoc k-NN implementation were used for each classifier respectively. Moreover, the general scheme for stacking has been modified to adapt it to geographic data. In this way, the first level (steps 5 and 6) consists in a SVM which takes every feature from the pixels in the training area to build an initial model which classifies every pixel from the study zone. At that point, a classical SVM application on images is resulted.

Input
l: LIDAR data
o: Orthophotograhy data
Output
m: LULC map
Begin
1. Develop a matrix *raster* in which every cell involves a physical position
2. Add the corresponding statistics from *l* and *o* to each pixel in *raster*

3. Select a training set from *raster*, called *train*
4. Label each pixel in *train* using expert knowledge
5. Build a SVM model, *svm*, from *train*
6. Use *svm* to classify every pixel in *raster*
7. For each pixel *p* in *raster*
   7.1. Collect the neigbourhood of *p* in a set *s*
   7.2. Build a k-NN model, *knn*, from *s*
   7.3. Use *knn* to classify *p*
8. Return a map *m* with every pixel spatial position and its label
End

Fig. 1. The LULC classification method based on a R-STACK algorithm (steps 6 to 8)

The novelty of the R-STACK method settles on the second level (step 7). Particularly, on the application of several classifiers (k-NN) and the way they are trained. Thus, a k-NN is build for each pixel taking its neighbours in the raster as training set which involves a strong relation (physical dependence) among the training pixels and the current pixel to classify. For the study area, we work with $k = 3$ and an 8-adjacency that is, each 3-NN is developed with just 8 instances of the pixel surrounding area. For this reason, the process can be tackled from the point of view of efficiency and complexity. In the end, the k-NN classifies the current pixel again having used the model built by its neighbours. In this way, possible inconsistences and non-desired effects can be removed. It is important to point out that it is necessary to make a raster copy before this last process and whilst the classes in the original raster are modified, every k-NN has to be build taking the neighbours from the raster copy in order to avoid collateral

Table 1. Sixty-one candidate variables. Variables with (*) are calculated for each band of a pixel: Height(H), Intensity(I), Red(R), Green(G) and Blue(B).

Variable	Description	Variable	Description
SNDVIMIN	SNDVI minimum	ICV	Intensity coefficient of variation
SNDVIMAX	SNDVI maximum	HCV	Height coefficient of variation
SNDVISTD	SNDVI Standard deviation	SLP	Slope
SNDVIAVG	SNDVI average	CRR	Canopy relief ratio
MIN(*)	Minimum	PEC	Penetration coefficient
MAX(*)	Maximum	TOTALR	Total of returns
STD(*)	Standard deviation	PCTN1	Unique return percentage
AVG(*)	Average	PCTN2	Double return percentage
VAR(*)	Variance	PCTN3	Three or more returns percentage
SKEW(*)	Skewness	PCTR1	First return percentage
KURT(*)	Kurtosis	PCTR2	Second return percentage
RANGE(*)	Range	PCTR3	Third or later return percentage
NOTFIRST	Second or later return	PCTR31	PCTR3 over PCTR1
EMP	Empty neighbours	PCTR21	PCTR2 over PCTR1
		PCTR32	PCTR3 over PCTR2

effects. Otherwise, the new classification sequence would affect the result of the remaining pixels.

4 Results

Two kinds of testing have been carried out to compare the efficiency of our approach against a classical SVM. The first test is based on statistical techniques. Since remote sensing data is expensive to generate, the comparison has to rest on an artificial data split. In our case, 100 splits are created from the original data so that each split contains about 740 instances. Then, a 10-fold-cross-validation process is made for every split. The results are registered for the following comparison process.

We have used the procedure suggested in several works [12] for robustly comparing classifiers across multiple datasets in order to evaluate the statistical significance of the measured differences in algorithm ranks. The chosen procedure involves the use of a statistical test to compare classifiers one each other. Our objective was to compare a classical SVM to our approach in terms of accuracy. Thus, the Wilcoxon procedure was selected as the appropriate test.

A fair comparison of the algorithms is obtained by average ranks and in this case, after the previous 100 10-fold-cross-validation results, our approach ranks first. With the measured average ranks, the Wilcoxon test checks whether the average ranks are significantly different from the mean rank r = 1.5 expected under the null hypothesis. Leaning on a statistical package (MATLAB), p value for the Wilcoxon test have resulted on a value less than $5.72e-06$ so the null

Table 2. A summing up of the hold-out test for the SVM classical approach

User class \sample	Water	Marsh	Roads or railways	Low Veg.	Middle Veg.	High Veg.	Buildings	Coal deposits	Dumps	Mixed areas
Water	178	5	0	0	0	0	0	0	0	2
Marshland	0	100	1	2	2	1	0	2	0	1
Roads or railways	0	4	69	0	6	0	0	0	1	0
Low Veg.	0	4	2	50	1	0	0	0	0	0
Middle Veg.	0	9	2	2	21	3	0	0	0	0
High Veg.	0	0	0	0	0	26	0	0	0	0
Buildings	0	2	3	0	2	1	31	0	0	0
Coal deposits	0	1	0	4	1	0	0	10	0	0
Dumps	1	0	0	0	0	0	0	0	21	9
Mixed areas	0	0	17	0	0	0	0	0	1	2
TP Rate	0.962	0.917	0.863	0.877	0.568	1.0	0.795	0.625	0.677	0.1
FP Rate	0.002	0.051	0.048	0.015	0.021	0.009	0	0.003	0.004	0.021
Precision	0.994	0.8	0.734	0.862	0.636	0.839	1	0.833	0.913	0.143
KIA	0.815									
Correctly classified	**0.846**									

Table 3. A summing up of the hold-out test for the SVM + k-NN restricted stacking

User class \sample	Water	Marsh	Roads or railways	Low Veg.	Middle Veg.	High Veg.	Buildings	Coal deposits	Dumps	Mixed areas
Water	181	3	0	0	0	0	0	0	0	1
Marshland	1	98	1	5	1	1	0	0	0	2
Roads or railways	0	4	72	0	2	0	0	0	0	2
Low Veg.	0	2	2	53	0	0	0	0	0	0
Middle Veg.	0	4	0	5	25	3	0	0	0	0
High Veg.	0	0	0	0	0	26	0	0	0	0
Buildings	0	2	2	0	2	2	31	0	0	0
Coal deposits	0	1	1	4	0	0	0	10	0	0
Dumps	1	0	0	0	0	0	0	0	24	6
Mixed areas	0	0	15	1	0	0	0	0	0	4
TP Rate	0.978	0.899	0.9	0.93	0.676	1.0	0.795	0.625	0.774	0.2
FP Rate	0.005	0.033	0.04	0.028	0.009	0.01	0	0	0	0.019
Precision	0.989	0.86	0.774	0.779	0.833	0.813	1	1	1	0.267
KIA	0.847									
Correctly classified	**0.873**									

hypothesis is rejected. Having found that the measured average ranks are significantly different (at $\alpha = 0.05$), our analysis based on ranks reveals that the accuracy of classical SVM is significantly worse than that of our approach for this kind of data.

The second type of testing is a hold-out process with data previously classified. This is the common testing in remote sensing. The test data set(600 instances) was selected from the original data set because of its special difficulty to be classified and it is not part of the training set. In Table 3 and Table 2, results for our approach and classic SVM are shown when the hold-out process is carried out. The general improvement is a 3% which is a very important advance.

5 Conclusions

In this paper, a new method based on a multiple classifiers ensemble was used to improve LULC map accuracy. The method built a statistical raster from LIDAR and image fusion data following a pixel-oriented strategy. Then, the pixels from a training area were used to train a SVM and k-NN restricted stacking (called R-STACK) taking into account the special characteristics of spatial data. A comparison between a SVM and the R-STACK method was carried out. The results in a riparian area of Huelva (Spain) showed a global accuracy of 84.6% for the classical SVM and 87.6% for the new approach which means a significant advance.

Even though results are satisfactory, there are still several problems to fix. Some of them are related to shadows from images and its weight on the final classification which has to be taken into account. Hence, a control of weights for each feature has to be implemented in order to avoid their misclassification effects. Genetic algorithms could be a very suitable tool to solve this problem. In addition, dependence on the training set can be a more important problem. Sometimes, the training set can be incomplete or not enough to describe the real space. These problems are harder to fix. Despite the fact that a semi-supervised approach seems to be more suitable to sort out this kind of problems, very few semi-supervised proposals can be found yet and more research is needed in order to develop them with the required accuracy. Finally, some problems are inherent in pixel-oriented approaches such as the detection of partial artificial structures. In the future, it would be very interesting to apply a prior phase in which at low addition to the computational cost, an object-oriented segmentation and classification could be carried out to extract the most difficult structures to classify, using visual recognition techniques from the computer vision world.

Acknowledgments. We would like to thank the Regional Ministry of Andalusia for all the support received in the development of this work and especially, to thank Irene Carpintero, Juan José Vales and Daniel Laguna for their very appreciated comments. We would also like to thank Francisco Martínez-Álvarez and Luis Gonçalves-Seco for all the time they invested that allowed this work to be completed.

References

1. Verrelst, J., Geerling, G., Sykora, K., Clevers, J.: Mapping of aggregated floodplain plant communities using image fusion of casi and lidar data. International Journal of Applied Earth Observation and Geoinformation (11), 83–94 (2009)
2. Antonarakis, A., Richards, K., Brasington, J.: Object-based land cover classification using airborne LIDAR. Remote Sensing of Environment (112), 2988–2998 (2008)
3. Dalponte, M., Bruzzone, L., Vescovo, L., Damiano, G.: The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. Remote Sensing of Environment (113), 2345–2355 (2009)
4. Shao, G., Wu, J.: On the accuracy of landscape pattern analysis using remote sensing data. Landscape Ecology (23), 505–511 (2008)
5. Fauvel, M., Benediktsson, J., Chanussot, J., Sveinsson, J.: Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. IEEE Transactions on Geoscience and Remote Sensing 46(11), 3804–3814 (2008)
6. Mitrakis, N., Topaloglou, C., Alexandridis, T., Theocharis, J., Zalidis, G.: Decision fusion of GA self-organizing neuro-fuzzy multilayered classifiers for land cover classification using textural and spectral features. IEEE Transactions on Geoscience and Remote Sensing 46(7), 2137–2152 (2008)
7. Waske, B., van der Linden, S.: Classifying multilevel imagery from sar and optical sensors by decision fusion. IEEE Transactions on Geoscience and Remote Sensing 46(5), 1457–1466 (2008)
8. Cortijo, F.J., Blanca, N.P.D.L.: Improving classical contextual classifications. International Journal of Remote Sensing 19(8) (1998)
9. Goncalves-Seco, L., Miranda, D., Crecente, R., Farto, J.: Digital terrain model generation using airborne LIDAR in florested area of Galicia. In: Proceedings of 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Spain, pp. 169–180 (2006)
10. Hudak, A.T., Crookston, N.L., Evans, J.S., Halls, D.E., Falkowski, M.J.: Nearest neighbor imputation of species-level, plot-scale forest structure attributes from lidar data. Remote Sensing of Environment 112, 2232–2245 (2008)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1) (2009)
12. Garcia, S., Herrera, F.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)